MPI Program Fails or Hangs

Category: Resolved

Problem: MPI program fails or hangs due to network communication problems.

Status: Resolved

Actions

Updated 02.06.12 - Our systems staff continue to replace bad or unreliable cables as they are detected. If you still experience this problem, contact the NAS Control Room: (800) 331-8737, (650) 604-4444, support@nas.nasa.gov.

Updated 06.20.11 - NAS systems staff are monitoring for errors in the InfiniBand fabric, and replacing bad or unreliable cables when detected.

Tips

If your MPI job aborts or hangs due to InfiniBand problems, your PBS output file will produce error messages similar to the below:

```
MPI ERROR: 14:34:10: rank 960: r199i0n2 IB board mlx4_0 port 1

had fault with communications to r190i0n6, restarting...
```

In this case, we recommended that you do the following:

1. Wait a few minutes and resubmit your job
2. File a ticket with the NAS Control Room to report the problem
3. If you have not done so, use SGI's MPT versions 2.0.4 or later, which are more robust against these InfiniBand problems and provide more diagnostic information in the system log files

Background

The network backbone of Pleiades comprises a pair of InfiniBand fabrics that are the currently the largest in the world (for details, see Network Resources). Most of the time, the large number of switches and cables works well, but sometimes, a cable will go bad, or its connection will work loose, causing some data packets to be lost or corrupted. When one cable fails, packets get re-routed, putting additional load on other paths, which can result in congestion and dropped packets.

The Lustre and TCP/IP protocols generally handle these failures by detecting bad or missing packets and retrying. The various MPI implementations cope less well, and with different degrees of success.

So, if your MPI program aborts with an error message that suggests some node had communication problems with another node (see above) or if the program hangs after issuing such an error message, then the program might have been affected by a cable failure.

Be aware that some communication errors are not caused by bad hardware - one rank running out of memory can cause communication error messages from surviving ranks.

While we monitor for errors in the InfiniBand fabric, and replace bad or unreliable cables when we detect them, paradoxically, the act of replacing a cable can cause its own errors. Lately, there has been an increase in cables needing work, possibly as a result of the recent facility over-heating incident in early April 2011.

Article ID: 277
Last updated: 30 Jul, 2012
Updates on Issues -> Resolved -> MPI Program Fails or Hangs
http://www.nas.nasa.gov/hecc/support/kb/entry/277/?ajax=1